Lernende Systeme
GERMANY'S PLATFORM FOR ARTIFICIAL INTELLIGENCE

Edge AI: AI close to the device

White Paper

Ecker, W., Houdeau, D. et al.
Working Group Technological Enablers and Data Science
Working Group IT Security, Privacy, Legal and Ethical Framework

Executive Summary

Edge AI (Edge Artificial Intelligence) refers to the deployment of artificial intelligence directly on or near devices – such as servers, vehicles, robots, smartphones, and even sensors. This enables real-time monitoring of health conditions or allows driver assistance systems to react instantly to obstacles, enhancing traffic safety. Edge AI fundamentally differs from Cloud AI: while Cloud AI processes data centrally using extensive computing and storage infrastructures, Edge AI processes data directly at or near the data source. This leads to numerous advantages, including lower latency, real-time applications, reduced energy consumption, and improved privacy and security. As such, Edge AI is a versatile and crucial technological foundation for innovative AI-driven solutions, especially in addressing 21st-century challenges like climate change, digital sovereignty, and energy management.

Deploying AI „on the edge“ presents certain challenges – limited computing power and storage capacity – but this shift brings (energy) efficiency and sustainability to the forefront of research, development, and application. Precisely because of the constrained processing power and storage of many devices, Edge AI has the potential to become a driver of resource-efficient AI innovations across various industries, markets, and fields of application.

Edge AI is emerging as a parallel trend to the energy-intensive reliance on central computing resources, commonly seen in generative AI today. Image generators like DALL-E or Stable Diffusion, for instance, require as much energy to create a single image as it takes to charge a smartphone. By designing Edge AI solutions explicitly for resource-constrained environments – often through smaller AI models – and leveraging methods like targeted data filtering and intelligent algorithm control, greater sustainability and energy efficiency can be achieved. For example, pacemakers store and analyze only relevant data, such as heartbeat anomalies, while AI systems in mobile devices, like electric vehicles or smartphones, can be selectively activated or deactivated based on energy availability.

Local processing also enables secure, reliable real-time data analysis, which is especially crucial for sensitive data, such as health records in medicine or valuable corporate information. However, this is only effective if the edge devices are adequately protected from unauthorised access.

Technical background

Various approaches can be employed to implement Edge AI (see Figure 1, options 1–7). AI training and AI inference can be carried out in different locations, including cloud servers, edge devices, or intermediate edge servers (also called fog). However, the distinction between AI inference and AI training isn't always clear-cut. In current practice, AI training typically demands significantly more storage and computational resources than AI inference. As a result, solutions where AI training occurs in the cloud while AI inference runs on edge devices are often more practical (see Figure 1, options 2–3).

Tackling challenges with Edge AI

The following advantages make Edge AI a valuable tool for addressing social and economic challenges.

- In smart grids, Edge AI technologies can forecast power generation and consumption, enabling efficient load management across the grid. In the realm of environmental sustainability, Edge AI can be integrated into recycling or waste treatment facilities using camera systems and sensors to identify and analyze various materials in real-time (see use case 'effective circular economy' by Plattform Lernende Systeme: Effektive Kreislaufwirtschaft: Roboter für eine bessere Trennung von Wertstoffen). With the help of actuators, these materials can be automatically sorted for specific recycling processes, increasing efficiency and supporting a circular economy. These are just a few of many possible Edge AI applications!

- Furthermore, this technology unlocks economic potential in areas such as production and logistics through automation and predictive maintenance, thereby enhancing competitiveness. Moreover, Edge AI can provide the necessary framework for companies that prefer not to share or externally process their sensitive raw data or lack sufficient data themselves to benefit from the rapid advancements in artificial intelligence.

- Additionally, Edge AI empowers organizations and businesses to operate with greater independence from predominantly non-European cloud-providers by processing data streams directly at the source, strengthening control over sensitive data. In this way, Edge AI can contribute to digital sovereignty.

Accelerating the translation of knowledge into applications

Germany is well-positioned to harness the potential of Edge AI, thanks to a blend of established technological expertise, domain knowledge, and experience in physical product development. The foundation for Edge AI success is already in place: Germany and Europe have considerable expertise, know-how, and excellent research in this field. Furthermore, in areas like mechanical engineering, machine vision, and sensor technology, Edge AI is already well advanced and established in terms of productive application.

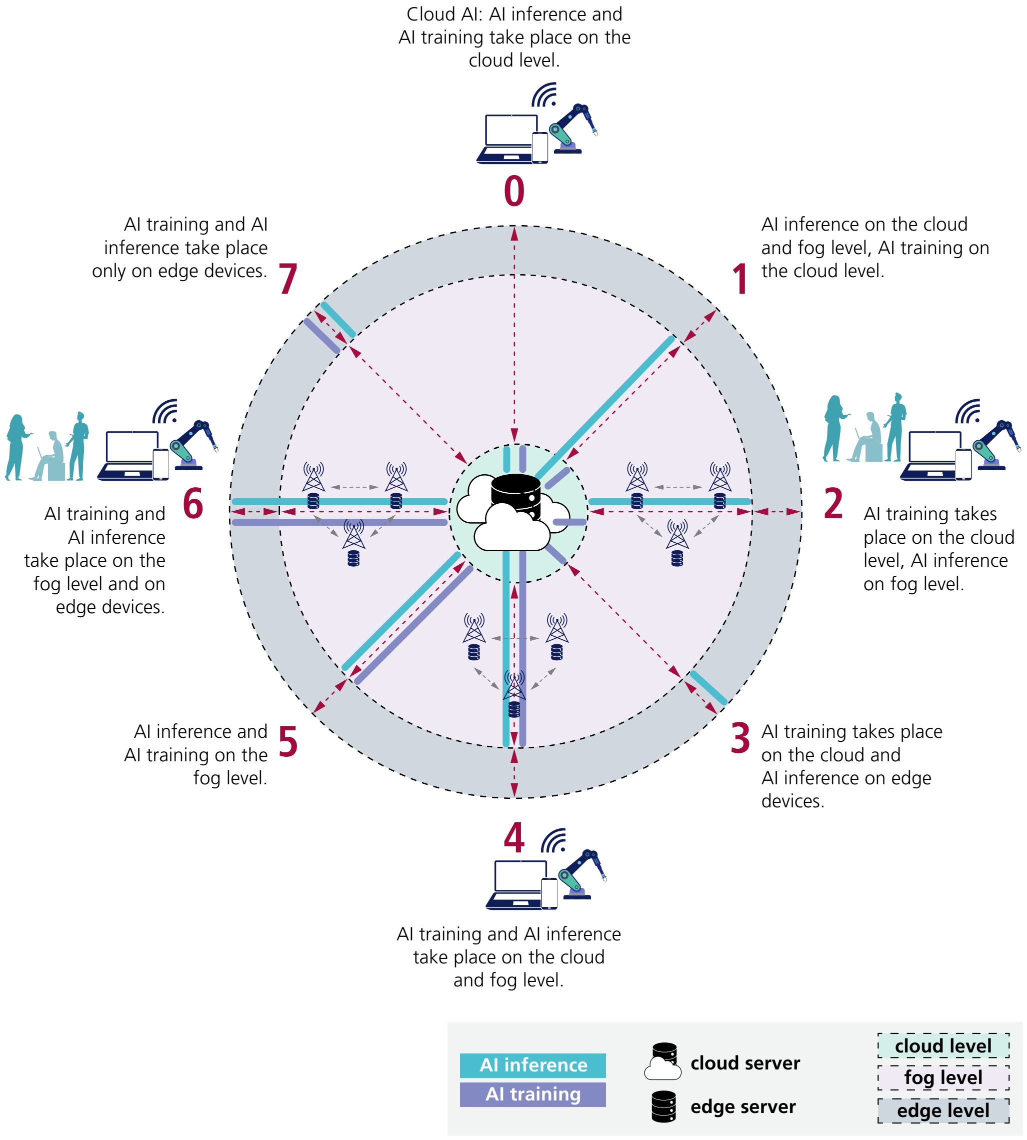
Proximity to key industries, including automotive, medical technology, and electronics, provides the opportunity to bring Edge AI technology into (widespread) use more quickly. Leveraging these diverse knowledge resources, companies in Germany can position themselves as problem solvers for complex technical challenges, with Edge AI as the key enabler.

However, there are several challenges to transferring these potentials into practice. Research and development face significant constraints due to the limited processing power and energy resources of smaller devices, alongside the need to keep costs manageable. Even when researchers identify a viable technical solution for Edge AI, building a complete system that meets all requirements remains challenging. This calls for an integrated approach: AI, software, and hardware must be co-developed to leverage and create synergies between them, achieving specific goals at the overall system level – a concept known as software-hardware co-design.

For instance, solutions must often be tailored to specific devices, such as ultra-low-power devices, or learning algorithms must be newly developed for particular hardware. Furthermore, the diversity of implementations, each with its unique program libraries, and the lack of standardized benchmarks make it difficult to translate research findings into real-world applications. Finally, sector-specific adoption presents an additional challenge in adapting Edge AI solutions across industries, as each field has distinct regulatory, certification, and compliance requirements.

To strengthen the transfer of Edge AI into practical applications, research, development, and industry should focus on creating platforms that offer foundational building blocks for Edge AI, adaptable to the needs of specific sectors. This approach requires the coordination of various stakeholders and standardization, particularly in terms of interfaces and hardware and development environments. Additionally, research should advance methods for resource-efficient data processing within the constraints of devices.

Different distributions of inference and training over various levels for Edge AI


Cloud AI: AI inference and AI training take place on the cloud level.
0
AI inference on the cloud and fog level, AI training on the cloud level.
1
AI training takes place on the cloud level, AI inference on fog level.
2
AI training takes place on the cloud and AI inference on edge devices.
3
4
AI training and AI inference take place on the cloud and fog level.
AI inference and AI training on the fog level.
5
AI training and AI inference take place on the fog level and on edge devices.
6
AI training and AI inference take place only on edge devices.
7
AI inference
AI training
cloud server
edge server
cloud level
fog level
edge level

Source: Custom representation based on McEnroe et al. (2022). Laptops, cell phones or on-board computers and other devices can also operate as edge servers (e.g. as edge servers for sensors).

Imprint
Editor: Plattform Lernende Systeme – Germany’s Platform for Artificial Intelligence | Managing Office | c/o acatech |
Karolinenplatz 4 | D-80333 Munich | kontakt@plattform-lernende-systeme.de | www.plattform-lernende-systeme.de |
Follow us on X: @Lernende Systeme | LinkedIn: de.linkedin.com/company/plattform-lernende-systeme |
Mastodon: social.bund.de/@LernendeSysteme | Status: July 2024 | Photo credit: gorodenkoff/iStock/Title

This executive summary is based on the white paper *KI nahe am Endgerät. Technologie für mehr Datenschutz, Energieeffizienz und Anwendungen in Echtzeit*, Munich, 2024. The authors are members of the working group Technological Enablers and Data Science and of the working group IT Security, Privacy, Legal and Ethical Framework. The original version of the publication is available online at: https://doi.org/10.48669/pls_2024-4

SPONSORED BY THE

Federal Ministry
of Education
and Research

acatech
NATIONAL ACADEMY OF SCIENCE AND ENGINEERING